STAT 302 Course Project 2 Instructions

1. **Using the glass dataset on Canvas** - determine the regression equation for predicting the response (Y) being variable (Si).

   Data description: https://archive.ics.uci.edu/ml/datasets/glass+identification

   - a) Select any three or four variables of your choice and use one visualization plot to tell a more complete and compelling story of the dataset masking use of all the variables selected. Consider using at least **color, faceting, theme** among many others in ggplot2. Write at least two paragraphs for this. One explaining why you choose those variables and the other what you see from the graph.
   - b) Fit **at least two different models** and select the best competing model for statistical analyses. Argue carefully on your choice of the best model. Perform optimization in R to compute the regression coefficients of your best model.
   - b) Use the results from your best model to answer the following questions:
     - i) What is the coefficient of determination and what does it mean?
     - ii) Find the least-squares estimates for the regression line.
     - iii) Interpret the value of the slopes and intercept in the context of the problem. In addition, state which variables are significant predictors at 10% level of significance? Explain.
     - iv) Perform a residual analysis to decide whether considering the assumptions for regression inferences met by the variables in the dataset appears reasonable.
   - c) We separate observations into two groups: `group 1: type=WinF` `group 2: type=WinNF`. Based on variable `Al`, are the distributions of the two groups significantly different under significance level $\alpha = 0.10$?
   - d) Perform a simulation analyses to micmic your best model. Explain your choose of values (it's completely up to you but should be reasonable values). Calculate the intercept and slopes for the least-squares regression line for the simulated data. Examine the residuals from the real data and the simulated data. Make sure to comment.

2. a) Using **Handout 1 dataset** on Canvas. Estimate the the prediction error and the prediction error rate of the model using either (i) Training and Testing or (ii) 5-fold cross-validation. The description of the dataset can be found on Canvas and select your own independent variables to use.

   b) Simulation is often used as a tool to teach statistics. There are many theoretical results that are easier to explain to students through a simulation. Choose a statistical concept, theoretical result, or problem from a homework or exam from a previous statistics class. **Design a simulation that would teach this topic or answer this question.**

   Some example topics include: simple random samples vs convenience samples, the probability of a certain event, the sampling distribution of the sample mean, confidence intervals, the central limit theorem, the difference between median and mode for data from a symmetric distribution vs a skewed distribution, etc. However, feel free to choose your own statistical topic! Please run your chosen topic by me in an email and get it approved before you start this part. Please **do not use Normal Distribution.**

   You should write this section as if you are teaching another student. Start with identifying the statistical topic and giving a brief overview. Then, run your simulation. Show your results in a table or plot. Explain your results to the student. Conclude with what you want them to take away from this simulation.

   c) Write a function to compute the mean, median, mode, variance, and skewness of Rayleigh distribution for (i) $\sigma = 6$; and (ii) $\sigma^2 = 10$). In addition, find the maximum likelihood estimator of σ and σ^2 from the Rayleigh distribution. A Rayleigh distribution with the correct formulas can be found here:

   `https://en.wikipedia.org/wiki/Rayleigh_distribution`

3. a) Using **Handout 1 dataset** on Canvas. Estimate the accuracy of the model using either (i) Training and Testing or (ii) 10-fold cross-validation. In this question, you can decide to treat commitment as binary using the median as cut-off (that is less than the median = 0, and 1 otherwise). The description of the dataset can be found on Canvas and select your own independent variables to use.

   b) FunToys is famous for three types of toys: Cars, Animals, and Robots. Each year, near the holiday season, it receives large bulk orders for these items. To meet these orders, FunToys operates three small toy-making factories, A, B and C.

   Factory A costs $1000 per day to operate, and can produce 30 cars, 20 animals and 30 robots per day. Factory B costs $2100 per day to operate, and can produce 40 cars, 50 animals and 10 robots per day. Factory C costs $1500 per day to operate, and can produce 50 cars, 40 animals and 15 robots per day.

   This Christmas, FunToys is required to deliver 5000 cars, 3000 animals and 2500 robots. You are tasked with finding out what is the most cost-efficient way to meet the order. In order to solve this problem, write down what you want to minimize and write a code in R to solve this optimisation problem. Report the optimal solution, and the value of the objective function at that solution. Interpret the solution: what do these numbers mean?

   c) Write a function to compute the mean, median, mode, and variance of Weibull distribution for $\lambda = 6$ and k = 1. In addition, find the maximum likelihood estimator of the parameters from the Weibull distribution. A Weibull distribution with the correct formulas can be found here:

   `https://en.wikipedia.org/wiki/Weibull_distribution`

Note the following:

- **Answer Question 1 and one other question. Two questions in total.**
- Only one person in the group should submit the course project. Please sign the form. Else, your project won't be graded.